DOCUMENT RESUME

ED 190 602 TM 800 394

AUTHOR Montague, William E.
TITLE A Common Flaw in Research Design: Inconsistency between Learning and Testing Requirements.
PUB DATE Apr 80
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Advance Organizers; *Correlation; *Educational Research; *Expectation; Higher Education; Intentional Learning; *Research Design; *Research Problems; Response Style (Tests); Test Items; Validity
IDENTIFIERS *Test Instructions

ABSTRACT

A number of examples are presented to illustrate a common flaw in the published research on learning, memory, and instruction. Experimental subjects--often college students--have certain expectations about the problems they will be asked to solve and about the questions that will appear on reading comprehension or recall tests; these expectations influence what is learned, noticed, and remembered during the experiment. If the experimental questions, then, are very different from the subjects' expectations, or if the instructions given to the subjects are ambiguous or misleading, poor performance will result. If the learners' expectancies are not consistent with the final testing, then interpretations based upon the research may be flawed. This author has noticed that about half of the papers and published studies of learning have ignored this problem, in spite of previous warnings by other researchers. Perhaps the subtlety of the problem in some situations allows it to be overlooked. (Specific examples are described for situations in which different experimental groups receive different task presentations, and in which they receive identical content but with different instructions. Recommendations for further research emphasize the importance of the subjects' role in an experiment). (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190602

AERA Paper April 15, 1980

APR 1980 RECEIVED
U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A common flaw in research design: Inconsistency between learning and testing requirements*

*1

William E. Montague
Navy Personnel R & D Center
San Diego, CA 92152

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
W. Montague
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The problem

The purpose of this paper is to point out and discuss what we consider to be a defect or flaw in many published experiments in the literature on learning, memory and instruction. The defect in the studies, which is often masked by theoretical speculation, goes undetected and undiscussed. It diverts attention from substantive issues and creates unproductive speculation. It is wasteful of limited research resources and of journal space.

Let me illustrate the problem by asking you to participate in an exercise that is similar to conditions used in some of the experiments under discussion.

PLEASE READ THE FOLLOWING PARAGRAPH CAREFULLY AT YOUR NORMAL SPEED. YOU WILL BE TESTED AFTER YOU FINISH READING.

> When the clumsy maid spilled prune juice on the royal throne, it left a distinct odor that filled the entire room. She was sure to lose her head

TM 800394

> for staining the king's favorite chair, so she covered it with a soft quilt, fumigated the hall with lilac spray and pretended she never had her mishap.

1. Answer the test question without rereading the paragraph. What words in the paragraph rhyme with: goose, tomb, sleigh?

Did that question surprise you? Was it consistent with what you thought you would be asked? Let's try it again.

PLEASE READ THE FOLLOWING PARAGRAPH CAREFULLY AT YOUR NORMAL SPEED. YOU WILL BE TESTED AFTER YOU FINISH READING.

> The cobra whipped its head out of the basket and eyed the mongoose. Instantly its smooth brown body grew taut, ready for an attack. Its beady eyes glared in fury, its forked tongue flickered out and back like flashes of black lightning. Suddenly the mongoose darted forward and the cobra struck at it viciously, but the blow missed. Instantly the battle was over as the mongoose seized the leathery throat of its enemy and held tight until the thrashing ceased.

2. Answer the test question without rereading the paragraph. How many pronouns are contained in the paragraph?

You were probably not as surprised by the type of question this time, but it is unlikely that the question you expected was consistent with the one asked. Perhaps you looked for rhymes, e.g., 'loose' rhymes with 'mongoose.' In experiments of this sort, in the absence of more specific instructions, subjects probably attend to the story line. Therefore, they can be surprised by questions asking about

the number of sentences or the number of articles or pronouns in the paragraph. They expect questions asking whether mongooses and cobras are natural enemies. When instructions are ambiguous, that is, when there are no prequestions, or advance organizers to guide them, subjects set themselves for the coming test. When there are cues contained in the material that they can detect they pay attention to them. For example, they are likely to notice the colorful adjectives in the last paragraph and expect that that aspect of the material might be part of the later questioning. If such was the case, they would not be surprised to receive a question about the vividness of the prose.

People in experiments are suspicious, They expect to be tested, and college students, who make up the bulk of the subject population, often are somewhat familiar with the experimental paradigms. They approach the learning task as if it is a problem to solve. The correspondence or consistency between how they approach the task and the questions on the later test determines their performance. Just to illustrate the point further, let me quote a subject in a recently published study which manipulated the complexity of adjunct questions (Yost, Avila, & Vexler, 1977):

> "The questions were so easy that I didn't trust you, and I had to reread each question several times to make sure I answered it correctly."

The paragraph exercises were intended to give you an

intuitive feel for the type of experiments we will be discussing and for the importance of the consistency between what you expected to be tested on and what you were tested on, in determining how adequately you were able to answer the questions. The fact that "expectancy," or "set" influences what information you learn from materials, what you see in a picture, what you remember from an experience, is not a new observation. Such observations are research classics.

In experiments in which conditions of learning, structure of the materials, or new methods of presentation are of interest, any inconsistency between what a subject or a student expects on a test and the test the experimenter has invented, results in relatively poor performance on that test by that subject. Any ambiguity in the instructions, or purposeful misdirection of subjects about how they will be tested produces inconsistency and leads to poorer performance than if a subject's expectancies, that is, what s/he attends to, and the test are consistent. Considerable care must be exercised to control the consistency between what subjects expect to be tested on and what they are tested on because explanations, interpretations, theories, etc., often depend on the results of that testing. If the learning sets/expectancies for some (groups) of subjects are inconsistent with the final testing, the interpretation may be in error. This is a flaw.

When the problem is presented this way it may seem so obvious that it appears to be naive. Let me assure you that it is not. The literature provides abundant examples of experiments which report results and generate theoretical AERA meeting when I was a critic/discussant for a paper session. Four of the five papers in that session ignored the relationship between subject set or expectancy and what was tested later. Similar problems exist in half of the papers sharing the paradigm that I critiqued during the present meeting. Recently we examined a number of journals published during the past few years and found many studies possessed this defect. For example, in the issues of the articles using the paradigm of interest here. Most of those studies about half ignored the problem and their interpretations, explanations, and discussions were flawed by the oversight. Many of these articles can be expected to stimulate additional, probably similarly defective research.

Although the problem seems to be a frequent one encountered in the published literature it must be remembered that only some studies are flawed in this way. Furthermore, this defect has been recognized by other at various times in the past (e.g., Anderson, 1970, Glynn and DiVesta, 1977, Rothkopf, 1976) It seems that few of us have paid serious attention to those reminders. Perhaps the subtlety of the problem in some situations allows it to be overlooked.

Let me attempt to concretize further what we consider

to be the flaw in the research. There are implicit task demands in any experiment where the subject is asked to read or learn materials to be tested later. In the absence of explicit instructions the subject selects what test to prepare for, e.g., read for meaning or story line, count the pronouns, look for ambiguous sentences or misspellings, etc.. Most experimental manipulations can be viewed as contributing to the task demands and influence what subjects study, how they process the material, i.e., what they look for, attend to or select from the material, in anticipation of later testing.

When experimental manipulations (perhaps based on some theoretical issue such as, near and far transfer, the role of vividness of material for remembering, the content of questions on learning, etc.) give hints to the subjects about how s/he will be tested they can be misleading... misleading in terms of the consistency between what a subject under one manipulation does and the testing. Many researchers ignore this aspect of their subjects' learning strategies and select a test strategy because of some theoretical bias. It appears that many researchers do not view the learner as an active, generative, participant in his own learning. This leads to experimental designs that confuse task demands, treatments, and testing.

Some examples taken from published research

In Table 1 two general experimental paradigms are presented to provide a framework for discussing specific studies.

Table 1

- Two Basic Paradigms -

	Learning Conditions	Testing Conditions
Type A	Different groups get different presentations (task demands are implicitly different) e.g., organization, structure, media, instructional strategies, advance organizers, (adjunct questions)	Consistency depends on degree to which what Ss do during learning matches testing
Type B	All groups get identical content, but different instructions (task demands are explicitly different) e.g., levels of processing, learning strategy, (adjunct questions)	In flawed studies there is usually no consideration of consistency

Type A studies are those in which different groups encounter different presentations or content, thereby creating different implicit task demands. Usually a single criterion test is used to assess performance There is no consideration of the consistency of the test with the task demands of the learning task. Falling into this category are

experiments that manipulate organization or structure of materials, presentation media, advance organizers, instructional innovations, and some adjunct question studies.

As an example, the cobra paragraph was taken from an experiment of mine (Montague and Carter, 1973) concerned with the effect of the vividness of descriptions on the recall of connected discourse. We found that recall of the vivid paragraphs was superior to their more subdued counterparts. Was this due the stimulation of some cognitive mechanism(s) or could it be intentional or attentional in response to inferred task demands? We don't know, and this sort of experiment is not likely to reveal the underlying determinants. Other examples, typically of more interest to educational psychologists, are studies in which some innovative instructional method is contrasted with others. For example, whether or not questions are asked at intervals during studying, and perhaps their type. Advance organizers have been of interest for years. Many of these studies ignore the expectancies of the subjects in anticipating the test which are the result of particular experimental manipulation.

Let me describe a specific experiment to point out the misleading nature of some research attempting to determine the effectiveness of different conditions of

learning or methods of presentation. An appropriately controls. An article in a recent Journal of Educational Psychology purported to study the effectiveness of an automated interactive method for teaching subjects to locate and label cities on a map, and learn other facts about various countries. Groups of subjects learning under the different conditions were compared using two types of tests: 1-locating cities and places on a map, and 2-recalling factual information about the country. Let's look at the learning conditions. All groups answered factual questions about a country during learning. We would not expect the groups to differ on the post-learning test on items such as these.

The three groups differed in the learning tasks relevant to locating places on the map. This is the point where the problem creeps in. One group learned under conditions where an automated system presented a map, a location started blinking, the student was to label the place, and corrective feedback was supplied by the system. The second group was given a labeled map and was questioned about the country by an automated system. The third group was given an unlabeled map and also answered questions about the places in the country.

It is important to notice that these conditions provide different tasks for the subjects--only that for the first group is consistent with the criterion test that asks subjects to put labels on the map. The results showed that the first group could label places on the map better than the others. The groups all performed about the same on the factual test.

The authors concluded that the students learned significantly more with the interactive map display. Did they? They outperformed them on the labeling criterion test, but not on the factual test. In fact, the claim was that the dynamic display characteristics of the automated map display was reponsible for the effect. This interpretation is suspect if we remember that the only group that was asked to label the map during learning was the first group. Their learning and testing conditions were perfectly consistent. It is unsurprising that they would be superior to groups where the learning and testing task requirements were inconsistent.

There is one final point to make about this example. The authors were (or are) enamored of an instructional method using a particular device. They created two problems in their study. First, they selected a task, labeling maps, which their system teaches as the criterion task. They then invented a couple of "strawmen"

or "do something else" tasks in which subjects in learning didn't label maps. Then, because the consistent group was superior on that task they conclude that their dynamic method is the cause. Perhaps we could think of several other ways to teach subjects to locate places or features on maps. Then we could compare different approaches appropriately, and make decisions about which one is preferred.

Type B (refer to Table 1) includes those studies in which all groups get identical content/presentations, but their instructions as to what to do with them differ. In some cases only a single criterion test is given, while in others multiple tests are used. However, in flawed studies there is no consideration of the consistency between the explicit task demands created by the different instructions and the testing. Into this category fall studies concerned with topics such as, levels of processing of text, and some of those using adjunct questions in learning text. We've put adjunct questions in both categories because their categorization depends on whether the questions are viewed as content manipulations or as instructions to subjects. In either case consistency can be a problem. A common occurrence in these studies involves the manipulation of task demands in learning with testing consistent with only one.

Examples of this sort of manipulation abound in the studies of levels of processing, and in many educational experiment counterparts. John Ellis will describe some of these studies in a few moments. But first, let me provide you with an advance organizer for his paper.

Several years ago a student of mine carried out a study in which different groups were given different tasks during learning and later tested for recall. They were shown stories like the first paragraph I showed you. One group then answered ten questions of the following form: The servant was a _____(maid); She dropped a glass of _____(juice). Another group found rhymes: suede (maid); goose (juice); for ten items. A third group was asked to find the word in the paragraph that filled the blanks: _a_d (maid); __ic_ (juice). The idea was to get subjects to process the material at different levels. The final test measured the ordered recall of the ten words. Notice that, in a sense, the subjects in the rhyme and spelling group were mislead. The test was not consistent with what they had been asked to do during learning. We did not measure recall of the rhyming words or how well subjects could search and find a word to fit the holes in a pattern.

Summary

It was my objective to focus attention on a problem common in studies which manipulate learning conditions, and either implicitly or explicitly what students expect on a later criterion test. The experimental flaw is that authors do not consider the possibility of inconsistency, and therefore, invent other explanations, often more complex, based on unnecessary or unwarranted assumptions, and misleading.

For the purposes of the discussions in this symposium studies have been categorized into two classes with somewhat different versions of the problem and needing different solutions. A few examples of such studies were discussed briefly. In the following presentations Wallace Wulfeck will discuss some of the category Type A experiments in more detail. Those are experiments in which implicit task demands pose potential inconsistency problems. Then John Ellis will discuss some studies falling into category Type B, involving learning conditions that evoke explicit task set which misleads some groups of subjects in terms of the criterion test used.

What we are proposing is a alternate (more parsimonious) theoretical view of the learner as a problem solver, a strong participant in determining what is learned in experiments and in instruction. Subjects adopt study strategies in anticipation of what the test will be like. Thomas

Anderson will discuss research that looks in more detail at the types of strategies subjects use when they have, or lack, information about the criterion testing. In addition, he will suggest several ways of solving or avoiding the problem.

A final observation provides additional perspective for the discussions to follow. First, it should be apparent that experimental reports should include more complete descriptions of the learning and testing tasks. In many of the studies we reviewed it was impossible to assess whether or not a consistency problem existed or not. We have some general prescriptions to assist in correcting the experimental flaws. For Type B experiments, different things are being learned because of the different instructions given to different groups. Criterion tests should evaluate these differences.

Experiments in category Type A present a tougher problem because of the more subtle nature of the implicit task demands. What seems to be required is more exacting analysis of both the learning and testing tasks to see what common processes there are. Such efforts might take the form of other experiments, or might involve taking protocols to identify subjects' expectancies. Some experiments clearly shouldn't have been done, either because authors set up arbitrary strawmen comparisons, or because the data can be explained in many equally unsatisfactory ways.

FOOTNOTES

*1. * The views expressed in this paper are those of the author and should not be construed as official views of the Department of the Navy or of the Federal Government. This paper was presented as the overview of the problem at a symposium at the Annual Meeting: American Educational Research Association, Boston MA, April 7-11, 1980.

References

Since the purpose of the paper is to focus attention on a problem common to many published studies in order to bring about some change, the particular studies (other than our own) are unlisted.

Anderson, R. C. Control of student mediating processes during verbal learning and instructioin. Review of Educational Research 1970, 40, 349-369.

Glynn, S.M. & DiVesta, F.J. Outline and hierarchical organization as aids for study and retrieval. Journal of Educational Psychology, 1977, 69, 89-95.

Montague, W.E. & Carter, J.F. Vividness in imagery in recalling connected discourse. Journal of Educational Psychology, 1973, 64, 72-75.

Rothkopf, E. Z. Writing to teach and reading to learn: A perspective on the psychology of written instruction. Ch. IV (pp 91-129) of the Seventy-fifth Yearbook The National Society for the Study of Education, Chicago, IL: NSSE, 1976.

Yost, M., Avila, L. & Vexler, E.B. Effect on learning of postinstructional responses to questions of differing degrees of complexity. Journal of Educational Psychology, 1977, 69, 399-408.